Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

Quantum Learning Enabled Green Communication for Next Generation Wireless Systems

Ankita Jaiswal, Sushil Kumar, *Senior Member IEEE*, Omprakash Kaiwartya, *Senior Member IEEE*, Pankaj Kumar Kashyap, Eiman Kanjo, Neeraj Kumar, *Senior Member IEEE*, Houbing Song, *Senior Member IEEE*

***Abstract*— Next generation wireless systems have witnessed significant R&D attention from academia and industries to enable wide range of applications for connected environment around us. The technical design of next generation wireless systems in terms of relay and transmit power control is very critical due to the ever-reducing size of these sensor enabled systems. The growing demand of computation capability in these systems for smart decision making further diversified the significance of relay and transmit power control. Towards harnessing the benefits of Quantum Reinforcement Leaning (QRL) in the design of next generation wireless systems, this paper presents a framework for joint optimal Relay and transmit Power Selection (QRL-RPS). In QRL-RPS, each sensor node learns using its present and past local state's knowledge to take optimal decision in relay and transmit power selection. Firstly, RPS problem is modelled as a Markov Decision Process (MDP), and then QRL optimization aspect of the MDP problem is formulated focusing on joint optimization of energy consumption and throughput as network utility. Secondly, a QRL-RPS algorithm is developed based on Grover's iteration to solve the MDP problem. The comparative performance evaluation attests the benefit of the proposed framework as compared to the state-of-the-art techniques.**

***Index Terms*– Quantum reinforcement learning, Wireless Systems, Internet of Things, Energy efficiency.**

I. INTRODUCTION

CONNECTED environment is growing significantly as next generation wireless systems around us [1]. The growing network of internet connected next generation wireless systems is termed Internet of Things (IoT) [2]. According to a recent report by Mckinsey, the number of connected IoT devices worldwide is projected to increase up to 43 billion by 2023 [3]. In the IoT enabled connected environment, balancing the usage of sensor's energy for longevity and maximizing the throughput of the sensor-enabled smart services is major concern [4, 5]. In next generation wireless systems, relay assisted wireless communication along with optimal transmit power has come out to be a favorable solution [6]. Nevertheless, proper relay allocation as forwarding sensors for data transmission is a crucial job as selection of relay might be led to lower throughput and higher energy consumption as compare to direct link transmission. Energy consumption can be significantly reduced by intelligently controlling the transmit power at IoT nodes [7].

A. Jaiswal, S. Kumar and P. Kashyap are with the Jawaharlal Nehru University, Delhi, 110067, India. Email:ankita79_scs@jnu.ac.in, skdohare@mail.jnu.ac.in, pankaj76_scs@jnu.ac.in

O. Kaiwartya and E. Kanjo are with Nottingham Trent University, NG11 8NS, UK, Email: Omprakash.kaiwartya@ntu.ac.uk, Eiman.Kanjo@ntu.ac.uk

N. Kumar is with the Department of computer science and information engineering, Asia University, Taiwan. Email: neeraj.kumar.in@ieee.org

H. Song is with the Embry-Riddle Aeronautical University, USA. Email: songh4@erau.edu

In the next generation stochastic wireless network, it is critical to have information about channel gain, energy harvesting capability, and neighboring systems interference in future for choosing an optimal relay and transmit power [8]. To solve the aforementioned problem, reinforcement learning (RL) approach has been used for optimal decision making [9]. For example, channel aware RL based Multi-path Adaptive routing has been suggested for optimal relay selection in WSN [10]. The RL based relay selection has been improved using Q-learning for better network performance in terms of reliability, outage probability, bit-error rate and network adaptivity [11]. Towards increasing learning rate in WSN, Deep-RL based relay selection scheme (DQ-RSS) has been investigated without utilizing prior network information [12].

Towards next generation wireless network design at "network-infrastructure and edge level" and "air interface and user-end levels" requires learning at user side, intelligent proactive caching, big data analytics, massive-IoT management, interoperability harmonization, distributed M-MIMO with fluid-antennas [13]. The use case for the next generation wireless network design consists of multi-dimension physical space, massive IoT device, High altitude platforms, body sensors, space shuttles etc. However, the classical RL technique is not suitable because of its slower learning performance, unexpected exploration and exploitation strategy and limited data analytics [14].

In this context, this paper presents a Quantum enabled RL framework for joint optimal Relay and transmit Power Selection (QRL-RPS). The framework focuses on harnessing the benefits of quantum computing enabled RL for design of next generation wireless systems in IoT centric smart service environment. The major contributions of paper are as follows:

1) Firstly, a system model is presented for sensors-enabled next generation IoT environment considering time varying energy harvesting and channel conditions.
2) Secondly, joint optimal relay and transmit power selection problem is formulated as a network utility maximization problem with energy, channel and data buffer constraints under classical RL. The MDP problem is transformed into QRL optimization problem using quantum representation.
3) Thirdly, a QRL-RPS algorithm is developed based on Grover's Iteration to solve the MDP problem. Convergence analysis and feasibility study of the proposed algorithm is done using temporal difference learning method and practical realization of quantum gates using quantum virtual machine, respectively.
4) The performance of the proposed algorithm is evaluated using Pyquil programming on Rigetti's Forest quantum virtual machine, and the results of the framework is compared with state-of-the-art techniques.

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

We organize the rest of the paper into the following sections. Section II reviews the related literatures. In section III, system model is presented. In section IV, RPS problem is formulated as an MDP. In section V, formulated maximization problem is represented using quantum reinforcement learning method. Section VI represents the proposed QRL-RPS algorithm to solve the formulated maximization problem and discusses its convergence property, time complexity and feasibility. In section VII, results of the simulation are discussed and analyzed. Finally, the conclusions and future scope of the paper is given in section VIII.

II. Related Work

A. Without Learning Approach

Several efforts in the research for optimal relay and transmit power selection have been made in recent years. In [15], a cognitive small world WSN and routing protocol have been suggested to balance energy and large data latency. In [16], communication protocol in WSN has been suggested for selection of an optimal relay from multiple candidate relays that has maximum value of harmonic mean function of its channel gains between source and relay, and relay and destination nodes. However, implementation of proposed algorithm has been done in static environment with non EH-nodes and does not explain the case of stochastic EH environment. In [17], authors have suggested a relay selection scheme. Here a relaying node is selected which has higher data transmission rate. However, the presented algorithm is prediction based which is not suitable for uncertain environment. In [18], authors have suggested optimal power allocation to access point and IoT devices in order to maximize energy efficiency of the full-duplex relay based IoT network. In [19-20], authors have suggested an adaptive online algorithm for joint relay selection, power allocation and time scheduling in order to optimize the system throughput. In most of the aforementioned research works, dynamic nature of environment is not considered. Further, it is assumed that perfect non-causal knowledge of energy arrival, channel fading and data arrival process in the network can be tracked down, which restricts its application to practical scenarios.

B. Learning Approach

Here, the focus has been given on utilizing RL methods for transmitting node to learn about the uncertain environment and then take optimal decision. In [11], authors have suggested a Q-learning based relay selection scheme (QL-RSA) that improves throughput, outage probability and bit-error rate of network. However, the above suggested schemes use a Q-table to save state-value and can only solve the problems with small states because the storage capacity of Q-table is limited. To sort out the Q-learning problem, a Deep RL in WSN was presented, named as DQ-RSS [12]. However, large number of relays and sensors results in high-energy consumption, which can cause node failure, but authors did not consider any method to provide continuous energy to nodes. An optimal policy based Bayesian RL for transmitting node to decide its data packet rate and transmit power has been suggested in [21]. However, the scheme is not suitable when state and action space are large enough. The convergence rate of scheme becomes slow for the complex problems. In [22], authors have suggested a novel approach based on deep RL to solve the problem of energy efficient power allocation. However, the proposed scheme has not considered the impact of time-variant nature of data arrival and energy arrival process. Nonetheless, no work has been done to jointly solve the problem of optimal relay and transmit power selection in sensor network via leaning method.

C. Quantum Reinforcement Learning

Recently, QRL has witnessed significant deliberation due to its advancement in computer vision. The concept of QRL is motivated by the concept of state superposition principle and quantum parallelism of quantum computing [23]. QRL has advantage over RL in improving the learning speed corresponds to balance between exploitation and exploration strategy [24]. Even though not many researches have been done in this field so far, some potential research has been done concerning QRL in the area of robot navigation [25]. However, QRL approach is yet unexplored in the field of sensor enabled IoT. Thus, motivated by the limitations of the existing works, the focus of our research work is to find an optimal policy for joint RPS, by exploring the idea of QRL in EH sensor enabled IoT network.

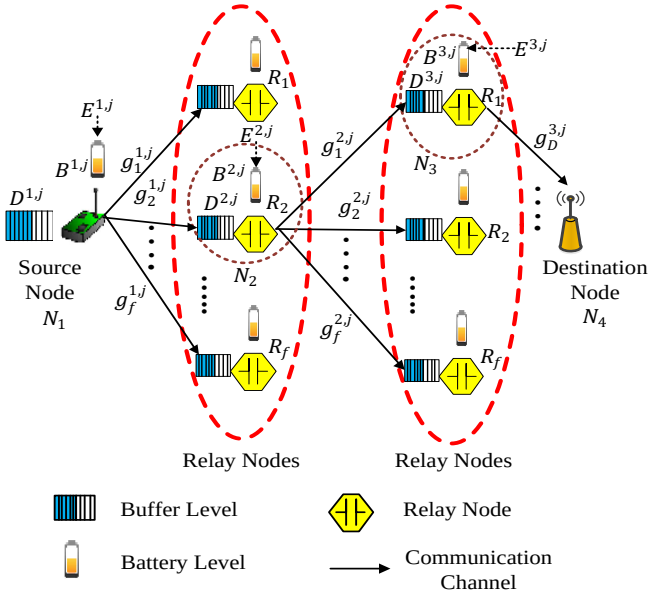

Fig. 1. Illustration of energy harvesting three-hop communication scenario

III. System Model

A. Network Model

We consider an EH sensors-enabled IoT system consisting of N nodes, F full duplex (decode and forward) EH relay nodes (R_f, for $f = 1,2,\ldots,F$) and a destination (sink) node, where nodes transfer data through three-hop relay communication model towards the destination node. The direct communication link between source and sink is assumed to be weak because of pathloss and fading problem. Therefore, the communication between source and sink is possible only via relays. These full-duplex relays are able to mitigate their self-interference and forward the data towards sink. For better understanding and simplicity of network, we

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

consider a scenario in which a single source N_1 transfers generated data to a destination N_4 through two optimally selected relays N_2 and N_3 as illustrated in the Fig. 1. At the beginning of each time slot t_j, source node N_1 generates and saves $R^{0,j}$ bits of data into its own buffer of size $D^{max,1}$. The relays N_2 and N_3 do not generate any data for transmission, these are only responsible for data transmission of other nodes. The relay N_3 and N_2 receive data from N_2 and N_1 respectively and store data into their finite buffer of size $D^{max,3}$ and $D^{max,2}$ respectively. The throughput of the links between N_1 and N_2, N_2 and N_3, N_3 and N_4 are limited by the arrival of data from N_1, N_2 and N_3 respectively. The objective of considered system is to maximize throughput of each link subject to minimum energy consumption.

Table 1. Frequently Used Notations

Notation	Description
R_f	f^{th} relay node
N_k	k^{th} sensor node
t_j	j^{th} time slot
ρ	Each time slot duration
$R^{0,j}$	Data bits generated by source node at time slot t_j
$R^{k,j}$	Throughput achieved at node N_k at time slot t_j
$D^{max,k}$	Maximum data buffer size of k^{th} sensor node
$E^{k,j}$	Amount of energy harvested by node N_k at time slot t_j
$E^{max,k}$	Upper bound on amount of energy harvested by node N_k
$B^{k,j}$	Battery level of node N_k at the beginning of t_j
$B^{max,k}$	Maximum battery capacity of node N_k
$\eta^{k,j}$	Energy harvesting efficiency of node N_k at time slot t_j
$\varepsilon^{k,j}$	Energy conversion efficiency of node N_k at time slot t_j
A^k_{solar}	Effective area of the photovoltaic cells mounted at the node N_k that absorb photons from solar energy
$E^{k,j}_{solar}$	Rate of absorbing photon from solar energy of photovoltaic cell mounted at the node N_k at time slot t_j
σ^2_k	Variance of additive white Gaussian noise at node N_k
$H^{k,j}$	Path loss coefficient vector for node N_k at time slot t_j
$h^{k,j}_F$	Path loss coefficient of transmission channel from node N_k towards R_f relay node at time slot t_j
D^k	Distance vector for node N_k
d^k_F	Distance between node N_k and R_f relay node
$M^{k,j}$	Multipath fading coefficient vector for node N_k at time slot t_j
$m^{k,j}_F$	Multipath fading coefficient value between node N_k and R_f relay
$G^{k,j}$	Channel gain vector for node N_k at time slot t_j
$g^{k,j}_F$	Channel gain value between node N_k and R_f relay node
$\mathfrak{P}^{k,j}$	Normalized radio propagation constant of node N_k at time slot t_j
α	Path loss exponent value
W	Channel Bandwidth
$P^{k,j}$	Optimal transmit power of transmitting node N_k at time slot t_j
$E^{k,j}_{con}$	Energy consumption of node N_k during time slot t_j
E^j_{con}	Total energy consumption during time slot t_j
$D^{k,j}$	Data buffer level of node N_k at time slot t_j
$U^{k,j}$	Utility of node N_k at time slot t_j

B. Energy Harvesting Model

Nodes and relay nodes harvest energy from ambient sources such as wind, solar, etc. Let node N_k, $k = \{1,2,3\}$ harvests $E^{k,j} \in \mathbb{R}^+$ $(joules)$ amount of energy in fixed time slot t_j, where $j = 1,2,\ldots.J$ and the length of each slot is constant and given by $\rho_j = t_{j+1} - t_j = \rho$. The node N_k can harvest maximum amount of energy $E^{max,k}$ depend upon energy harvesting source. The harvested energy is stored into battery of maximum capacity $B^{max,k}$. It is presumed that sensors' batteries are ideal in nature that means there is no energy loss during retrieving and storing of energy. As a special case, we consider solar power energy harvesting source. Let $\eta^{k,j} \epsilon (0,1)$ be energy harvesting efficiency, $\varepsilon^{k,j} \epsilon (0,1)$ be energy conversion efficiency and $A^k_{solar}(m^2)$ be effective area of photovoltaic cells mounted at nodes that absorb photons from solar energy at the rate of $E^{k,j}_{solar}(joule/m^2)$. Thus, the energy harvested $E^{k,j}(joule)$ by a node at time slot t_j is computed as

$$E^{k,j} = E^{k,j}_{solar} \times A^k_{solar} \times \varepsilon^{k,j} \times \eta^{k,j} \quad (1)$$

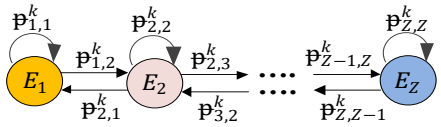

Fig. 2. Markov chain based EH model of the network with Z states

The amount of energy harvested by node varies over time. Thus, energy harvesting can be quantized to Z levels with $E^{k,j} \epsilon \{E_e\}_{1 \le e \le Z}$, $k = 1,2,3$; which are modelled as Markov chain accompanied by Z states as shown in Fig. 2 . The transition probability of $E^{k,j}$ from E_1 *to* E_2 during time slot t_j is given as $\mathfrak{p}^{k,j}_{1,2} = prob\,(E^{k,j} = E_2 | E^{k,j-1} = E_1)$.

C. Channel Model

The channel from transmitting node to receiving node is assumed to be quasi-static block Rayleigh fading model that means channel gain is different for different time slot but constant within time slot. The channel is assumed to be noisy and channel state information is known at the transmitting node through feedback from receiving node. Additive white Gaussian noise at N_k, with zero mean and variance $\sigma^2_k = \sigma^2$ introduce impairment into the channel that results in degradation in the strength of transmitted signal. Hence, intended receiver is not able to successfully decode the received information.

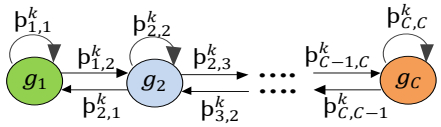

Fig. 3. Markov chain based channel model of the network with C states

Let $H^{k,j} = \left[h^{k,j}_1, h^{k,j}_2 \ldots h^{k,j}_F\right] \epsilon \mathbb{C}$ be path loss coefficient vector, $D^k = [d^k_1, d^k_2 \ldots d^k_F]$ be distance vector and $M^{k,j} = \left[m^{k,j}_1, m^{k,j}_2 \ldots m^{k,j}_F\right] \epsilon \mathbb{C}$ be multipath fading coefficient vector between transmitting and receiving node at t_j. Thus, channel gain vector $G^{k,j} = \left[g^{k,j}_1, g^{k,j}_2 \ldots g^{k,j}_F\right] \epsilon \mathbb{C}$ is computed as $G^{k,j} = \left|H^{k,j}\right|^2 \left|M^{k,j}\right|^2$, where $H^{k,j} = \mathfrak{P}^{k,j}(D^k)^{-\alpha}$, such that $\mathfrak{P}^{k,j}$ is normalized radio propagation constant and α is path loss exponent. High channel gain value improves network throughput. Further, the channel state is time-variant due to dynamic nature of environment. Thus, channel state can be quantized to C levels with $g^{k,j}_F \epsilon \{g_i\}_{1 \le i \le C}$, $k = 1,2,3$; which are modelled as Markov chain accompanied by C states as shown in Fig. 3. The transition probability of $g^{k,j}_F$ from g_1 *to* g_2 during time slot t_j is given as $\mathfrak{p}^{k,j}_{1,2} = prob\,(g^{k,j}_F = g_2 | g^{k,j-1}_F = g_1)$.

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

IV. Joint Optimal Relay and Transmit Power Selection Problem as Markov Decision Process

In this section, we aim to formulate the joint optimal relay and transmit power selection problem as Markov decision process (MDP) to represent the network utility maximization that jointly optimizes energy consumption and throughput. The amount of data (bits) successfully received at N_4 during t_j is defined as network throughput for respective time slot. The total power used for data transmission from source to destination is considered as total energy consumption for time slot t_j, because energy used for data sensing and receiving operation is very less as compared to transmission of data. As, there is no direct communication between source (N_1) and sink (N_4), the throughput $R^{3,j}$ at N_4 corresponds to the amount of data received by it from relay N_3 at t_j. The nodes N_3 and N_2 only send data that they have received from N_2 ($R^{2,j}$) and N_1 ($R^{1,j}$) respectively. Thus, $R^{3,j}$ is restricted by throughput $R^{2,j}$ and $R^{1,j}$. The throughput $R^{1,j}, R^{2,j}$ and $R^{3,j}$ obtained at time slot t_j are computed as

$$R^{k,j} = \rho W \log_2\left(1 + \frac{\left|g_f^{k,j}\right|^2 P^{k,j}}{\sigma^2}\right), \text{ for } \{f = 1,2,\dots F; k = 1,2,3, and\ j = 1,2,\dots J\} \quad (2)$$

where, $P^{k,j}$ is the allocated optimal transmit power of node N_k at time slot t_j, W is the channel bandwidth and $g_f^{k,j}$ denotes channel gain value between node $\mathrm{N_k}$ and $\mathrm{R_f}$ relay node at t_j.

Amount of energy available at N_k for data transmission is controlled by correspondent energy harvesting process. Transmit power allocation at N_k can be done only when the harvested energy has already been saved in battery. Here it is presumed that batteries can't be recharged immediately. As a result, at t_j battery only stores energy that has been harvested up to t_{j-1}. Consequently, there exists a causality restriction on energy consumption that needs to be accomplished,

$$E_{con}^{k,j} = \rho P^{k,j} \le B^{k,j}, k = \{1,2,3\} \quad (3)$$

Where, $B^{k,j}$ is the battery level of node N_k at the beginning of t_j and can be given as

$$B^{k,j} = \min\{B^{max,k}, B^{k,j-1} - \rho P^{k,j-1} + E^{k,j-1}\}, \ \{k = 1,2,3, and\ j = 1,2,\dots J\} \quad (4)$$

The battery capacity needs to be considered finite so that overflow condition could be avoided in which some part of harvested energy is wasted due to battery full state. Thus, the constraint on energy overflow is given as

$$B^{k,j} - \rho P^{k,j} + E^{k,j} \le B^{max,k}, k = \{1,2,3\} \quad (5)$$

As previously stated, data $R^{0,j}\ bits$ that are generated at N_1 during time slot t_j is an independent process. The data received at N_2 and N_3 depend on $R^{1,j}$ and $R^{2,j}$ respectively. Therefore, data buffer level $D^{k,j}$ of node N_k is computed as

$$D^{k,j} = \sum_{n=1}^{j-1} R^{k-1,n} - \sum_{n=1}^{j-1} R^{k,n}, \{k = 1,2,3\} \quad (6)$$

Further, the throughput $R^{1,j}$, $R^{2,j}$ and $R^{3,j}$ need to satisfy data causality constraints which make sure that N_k doesn't retransmit information that has not received yet and given as

$$R^{k,j} \le D^{k,j}, \{k = 1,2,3\} \quad (7)$$

Similar to the constraint on energy in (5), N_k has constraint on data buffer overflow given as

$$D^{k,j} - R^{k,j} + R^{k-1,j} \le D^{max,k} \quad (8)$$

Further, total energy consumption during t_j is computed as

$$E_{con}^{j} = \sum_{k=1}^{3} \rho P^{k,j} \quad (9)$$

Additionally, the presented framework can also be used for any number of hop communication, where each node consists of an independent EH process and channel condition. So, it can be solved as independent point-to-point communication problem, when only local causal knowledge is known at source and relays. Thus, local solutions for joint optimal relay and transmit power selection problem at each point-to-point communication leads to globally optimal solution for network. Subsequently, three point-to-point EH communication problem presented in fig .1 correspond to the links $N_1 \to N_2$, $N_2 \to N_3$ and $N_3 \to N_4$. Notwithstanding, as $R^{3,j}$ is restricted by throughput $R^{2,j}$ and $R^{1,j}$, the joint relay and transmit power selection problem for N_1, N_2 and N_3 is coupled. Thus, the utility maximization problem for node N_k to choose its optimal relay $R_f^{k,j}$ and transmit power $P^{k,j}$ for transmission of data towards sink during time slot t_j independently computed as

$$U^{k,j} = \ln\left(\frac{R^{k,j}}{E_{con}^{k,j}}\right) \quad (10)$$

The logarithmic function's concavity captures network utility better in terms of throughput and energy consumption. However, solution of above coupled utility maximization problem cannot be achieved by using only local causal knowledge, because EH-IoT nodes do not have information of other nodes regrading power allocation policy, data buffer, energy harvesting and channel. It means transmitter can't minimize data buffer overflow of receiving node by reducing its own transmit power. To find the solution of convex optimization problem, we formulate each point to point communication problem as an MDP due to: (i) MDP provide decision maker (learning agent) that interact with environment and take decision in control manner to achieve a goal (ii) MDP is able to handle the stochastic nature of EH and channel fading (iii) MDP estimates future utility using only casual knowledge of current state and taken action, it does not require non-casual knowledge in advance (iv) MDP build mathematical formula to efficiently use the QRL approach to find an optimal policy for selection of joint relay and transmit power.

Each node $N_k(k = 1, 2, 3)$ consists of a learning agent and its MDP comprises the following: (i) a set of possible states $S^k = B^k \times E^k \times D^k \times G^k$. (ii) A set of possible actions $A^k = {R_f}^k \times P^k$. (iii) A state transition function $T^k: S^k \times A^k \times S^k \to [0,1]$ that precisely identify the unpredictability about future transition amid states. (iv) A reward function $U^k: A^k \times S^k \times U^k \to [0,1]$. During each time slot t_j, the state $s^{k,j} \epsilon\, S^k = [B^{k,j} \times E^{k,j} \times D^{k,j} \times G^{k,j}]$ of node N_k includes current battery level, amount of harvested energy, data buffer level, and channel states to F relays that takes any value in continuous range. The action set $a^{k,j}(s^{k,j})\, \epsilon A^k = ({R_f}^{k,j}, P^{k,j})$ for corresponding state $s^{k,j}$ consists of all candidate relays $\{{R_f}^k, f = 1,2,\dots,F\}$ in the range of N_k and transmit power values $P^{k,j} = \{\delta^k, 2\delta^k, \dots B^{max,k}\}$ that N_k

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

selects, and δ^k is the step size. The state transition function refers to transition probabilities from $s^{k,j}$ to $s^{k,j+1}$ when an action $a^{k,j}(s^{k,j})$ is taken. The reward function $U^{k,j}$ is defined as the immediate reward in terms of utility obtained in time duration ρ for taking an action $a^{k,j}(s^{k,j})$ under state $s^{k,j}$. The utility $U^{k,j}$ is only known at N_k (transmitter) at the end of time slot t_j i.e, at t_{j+1} which is computed using throughput received as a feedback from the receiver and its energy consumption, to assess the quality of selected action. The main objective of learning agent at N_k is to obtain an optimal policy, $\pi^k(s^{k,j}) \rightarrow ({R_f}^{k,j}, P^{k,j})$ and it is a corresponding solution to the MDP problem. Further, $\pi^k(s^{k,j})$ can be assessed using state-value function $V^{\pi}(s^{k,j})$, which is illustrated as the expected reward starting from state $s^{k,j}$, choosing $({R_f}^{k,j}, P^{k,j})$ and following π^k afterwards. The optimal policy π^{k^*} consists of state value that is greater than or equal to state value of other policies for each state. The correspondent state-value for π^{k^*} is represented as $V^{\pi^*}(s^{k,j})$. Therefore, the expected sum of discounted reward for each state is given as

$$V^{\pi}(s^{k,j}) = E\{U^{k,j+1} + \gamma U^{k,j+2} + \cdots \left| s^{k,j}, \pi^k \right\} = E\{U^{k,j+1} + \gamma V^{\pi}(s^{k,j+1}) + \cdots | s^{k,j}, \pi^k\} \quad (11)$$

And, the temporal difference (TD) one step updating rule [26] of $V(s^{k,j})$ is given as

$$V(s^{k,j}) \leftarrow \mathrm{V}(s^{k,j}) + \zeta(U^{k,j} + \gamma \mathrm{V}(s^{k,j+1}) - V(s^{k,j})) \quad (12)$$

where, $\zeta \in [0,1]$ represent the learning rate, $\gamma \in [0,1]$ is discount factor and $U^{k,j}$ is obtained reward for corresponding action.

As the repercussion of only casual knowledge at transmitter, there exit a trade-off between power allocation in the current time slot to avoid battery and buffer overflow and energy is stored for next time slot might for bad channel condition. To consider this uncertainty, we prefer to maximize the network utility in current time slot over future time slot. Besides, a discount factor $0 \leq \gamma \leq 1$ is included to weight the higher network utility in current time slot versus attain higher network utility in future. As, γ close to zero give priority to current time slot and γ close to unity future time slot consider for utility maximization. In turn, the objective function (Eq. 10) i.e., joint optimal relay and transmit power allocation problem for network utility maximization of three links $\mathrm{N}_1 \rightarrow \mathrm{N}_2$, $\mathrm{N}_2 \rightarrow \mathrm{N}_3$ and $\mathrm{N}_3 \rightarrow \mathrm{N}_4$ is replaced by expected utility as the novelty of problem can be expressed as

$$(\mathrm{R}^{\mathrm{k,j}}_{\mathrm{f,opt}}, \mathrm{P}^{\mathrm{k,j}}_{\mathrm{opt}}) = \underset{\{\mathrm{R}^{\mathrm{k,j}}_{\mathrm{f}}, \mathrm{P}^{\mathrm{k,j}}\}}{\operatorname{argmax}} \lim_{\mathrm{J}\to\infty} \mathbb{E}\left[\sum_{\mathrm{j}=1}^{\mathrm{J}} \gamma^{\mathrm{j}} \mathrm{U}^{\mathrm{k,j}}\right] \quad (13)$$

Subject to- C1: $\sum_{\mathrm{j}=1}^{\mathrm{J}} \rho \mathrm{P}^{\mathrm{k,j}} \leq \sum_{\mathrm{j}=1}^{\mathrm{J}-1} \mathrm{E}^{\mathrm{k,j}}, \forall \mathrm{k}, \mathrm{j} = 1,2,..\mathrm{J}$

C2: $\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{E}^{\mathrm{k,j}} - \sum_{\mathrm{j}=1}^{\mathrm{J}} \rho \mathrm{P}^{\mathrm{k,j}} \leq \mathrm{B}^{\max,\mathrm{k}}, \forall \mathrm{k}, \mathrm{j} = 1,2,..\mathrm{J}$

C3: $\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{R}^{\mathrm{k,j}} \leq \sum_{\mathrm{j}=1}^{\mathrm{J}-1} \mathrm{R}^{\mathrm{k}-1,\mathrm{j}}, \forall \mathrm{k}, \mathrm{j} = 1,2,..\mathrm{J}$

C4: $\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{R}^{\mathrm{k}-1,\mathrm{j}} - \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{R}^{\mathrm{k,j}} \leq \mathrm{D}^{\max,\mathrm{k}}, \forall \mathrm{k}, \mathrm{j} = 1,2,..\mathrm{J}$

C5: $\mathrm{P}_{\min} \leq \mathrm{P}^{\mathrm{k,j}} \leq \mathrm{P}_{\max}, \forall \mathrm{k}, \mathrm{j} = 1,2,..\mathrm{J}$

Where C1 indicates constraints on transmit power over each time slot (ρ) during $\mathrm{t_j}$. C2 indicates constraint on energy overflow condition after power allocated to nodes; C3 specifies restriction on data rate. C4 indicates constraints on data buffer overflow. C5 indicates boundary condition constraint on allocated power $\mathrm{P}^{\mathrm{k,j}}$, value of lower bound $\mathrm{P}_{\min}$ and upper bound $\mathrm{P}_{\max}$ controlled by power amplifier's hardware circuit corresponds to Eq. (3), (5), (7), (8) and (9) respectively. The formulated problem in (13) can be optimally solved by QRL method motivated by the concepts of quantum mechanics.

V. Quantum Optimization Aspect of MDP Problem

The quantum computation in QRL is related to state superposition and quantum parallelism. Quantum bit (qubit) is the key to quantum computation similar to classical bits used in traditional computation. A single qubit consists of two basic states denoted by $|0\rangle$ and $|1\rangle$ (Dirac representation), corresponding to states 0 and 1. Moreover, qubit also exits in superposition state of $|0\rangle$ and $|1\rangle$. Thus, state of a qubit $|\Psi\rangle$ can be represented as linear combination of $|0\rangle$ and $|1\rangle$ defined as state superposition principle.

$$|\Psi\rangle = €|0\rangle + ¥|1\rangle \quad (14)$$

where, € and ¥ are complex probability amplitude of state $|0\rangle$ and $|1\rangle$ with magnitude (occurrence probability) $|€|^2$ and $|¥|^2$ respectively, satisfying the sum of probability condition $|€|^2 + |¥|^2 = 1$, and $|0\rangle$ and $|1\rangle$ form a set of orthonormal bases in Hilbert space. Applying, unitary transformation on superposition state of qubits simultaneously transforms all states (basis vectors) and provides new superposition state with different values known as quantum parallelism. Quantum gates such as Hadamard gate and phase gate are used to achieve quantum parallelism. To represent states of network system, we need multiple qubits. Let there are 2^u states then we require u-qubits quantum gates for computation. The quantum state $|\Psi_u\rangle$ of u qubits (base-10) denoted as linear combination of orthonormal states given as

$$|\Psi_u\rangle = \sum_{i=0}^{2^u-1} ¥_i |i\rangle = ¥_0 \overbrace{|000\ldots0\rangle}^{u} + ¥_1 \overbrace{|000\ldots1\rangle}^{u} + \cdots + ¥_{2^u-1} \overbrace{|111\ldots1\rangle}^{u} \text{ And, } \sum_{i=0}^{2^u-1} |¥_i|^2 = 1 \quad (15)$$

The advantage of QRL over traditional RL carried out by three strategies, (i) exploration policy in QRL depends upon collapse postulate rather than greedy or Boltzmann policy [26], which provides better balance between exploration and exploitation of available actions. (ii) QRL updates all the states simultaneously using unitary transformation method in synchronized manner. (iii) QRL algorithm is robust in nature because it adapts different learning rates and discount factor for unknown environment quickly. The robustness characteristic turnout as learning performance for the QRL algorithm as verified in the simulation section VII-B -3, 4.

A. Quantum State and Action Representation

QRL boosts the action selection probability and reinforcement strategy of classical RL algorithm by adopting the idea of Quantum computation. Collapse postulate and amplitude amplification are the policy of quantum computation used for selection of probabilistic action and action reinforcement respectively. The orthogonal quantum state $|s_l^{k,j}\rangle$ and its corresponding action $|a_w^{k,j}\rangle$ (Eigen state and Eigen action respectively) of node N_k at t_j in QRL are the quantized representation of same state ($s_l^{k,j}$) and action ($a_w^{k,j}$) in RL. The total number of discrete states can be written as a set $S^k = \{|s_1^k\rangle, |s_2^k\rangle, |s_3^k\rangle \ldots\ldots |s_l^k\rangle \ldots\}$ and corresponding

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

action set for state ($|s_l^k\rangle$) is given as $A^k(s_l^k) = \{|a_1^k\rangle, |a_2^k\rangle, |a_3^k\rangle \dots, |a_w^k\rangle \dots\}$. Let N_{st} and N_{ac} be the number of Eigen states and Eigen actions of a node, then we choose u and v number of qubits to represent them, which satisfy the following inequalities

$$N_{st} \le 2^u \le 2N_{st}\ ,\ N_{ac} \le 2^v \le 2N_{ac} \quad (16)$$

The qubit representation of a state and its corresponding action during t_j time slot is given as

$$\left|s_l^{k,j}\right\rangle = \left|B^{k,j} \times E^{k,j} \times D^{k,j} \times G^{k,j}\right\rangle = \{q_1, q_2, q_3 \dots . q_u\} \quad (17)$$

$$\left|a_w^{k,j}(s_l^{k,j})\right\rangle = \left|R_f{}^{k,j}, P^{k,j}\right\rangle = \{q_1, q_2, q_3 \dots . q_v\} \quad (18)$$

To explore QRL, Eigen states and its corresponding Eigen actions are represented in superposition state and action respectively. This superposition state, ($\left|s_l^{k,j^{N_{st}}}\right\rangle$) (action, $\left|a_w^{k,j^{N_{ac}}}(s_l^{k,j})\right\rangle$) is the sum of all existing quantum states (actions) in the system.

$$\left|s_l^{k,j^{N_{st}}}\right\rangle = \sum_{l=1}^{N_{st}} L_l \left|s_l^{k,j}\right\rangle \leftrightarrow \left|s_l^{k,j^u}\right\rangle = \sum_{s_l^k=00\dots0}^{\overbrace{11\dots1}^{u}} L_{s_l^k} \left|s_l^{k,j}\right\rangle \quad (19)$$

$$\left|a_w^{k,j^{N_{ac}}}(s_l^{k,j})\right\rangle = \sum_{w=1}^{N_{ac}} L_w \left|a_w^{k,j}\right\rangle \leftrightarrow \left|a_w^{k,j^v}(s_l^{k,j})\right\rangle = \sum_{a_w^k=00\dots0}^{\overbrace{11\dots1}^{v}} L_{a_w^k} \left|a_w^{k,j}\right\rangle \quad (20)$$

The probability amplitude of Eigen state $L_{s_l^k}$ and corresponding Eigen action $L_{a_w^k}$ are represented as complex number. When superposition state $\left|s_l^{k,j^{N_{st}}}\right\rangle$ is measured, it collapses into $\left|s_l^{k,j}\right\rangle$ with probability $\left|L_{s_l^k}\right|^2$. Also valid for measuring an action $\left|a_w^{k,j^{N_{ac}}}(s_l^{k,j})\right\rangle$, collapse into $\left|a_w^{k,j}\right\rangle$ with probability of $\left|L_{a_w^k}\right|^2$. The $L_{s_l^k}$ and $L_{a_w^k}$ satisfy the basic probability condition

$$\sum_{s_l^k=00\dots0}^{\overbrace{11\dots1}^{u}} \left|L_{s_l^k}\right|^2 = 1, \quad \text{and} \quad \sum_{a_w^k=00\dots0}^{\overbrace{11\dots1}^{v}} \left|L_{a_w^k}\right|^2 = 1 \quad (21)$$

B. QRL based Action Selection Policy

The probability of selecting an action in current Eigen state depends on collapse postulate of quantum computation. Collapse postulate is defined as when an action $\left|a_w^{k,j^v}(s_l^{k,j})\right\rangle = \sum_{a_w^k=00\dots0}^{\overbrace{11\dots1}^{v}} L_{a_w^k} \left|a_w^{k,j}\right\rangle$ is measured, it will be randomly selected and collapse into one of its Eigen action $\left|a_w^{k,j}\right\rangle$ and action state changes with corresponding probability ($\left|\langle a_w^{k,j} \left|a_w^{k,j^v}(s_l^{k,j})\right\rangle\right|^2$) .

$$\left(\left|\langle a_w^{k,j} \left|a_w^{k,j^v}(s_l^{k,j})\right\rangle\right|^2\right) = \left|(|a_w^{k,j}\rangle)^* \left|a_w^{k,j^v}(s_l^{k,j})\right\rangle\right|^2 = |(|a_w^{k,j}\rangle)^* \sum_w L_{a_w^k} |a_w^{k,j}\rangle|^2 = \left|L_{a_w^k}\right|^2 \quad (22)$$

In QRL, the agent selects an action according to sum of expected discounted reward at each state. The agent learns action selection policy π: $S^k \times \cup_{s_l^{k,j} \in S^k} A^k(s_l^k) \to [0,1]$, which maximizes the expected reward by updating state value. i.e., it is a mapping from state to action $f(s_l^k) = \pi$: $S^k \to A^k$.

$$f(s_l^k) = \frac{|a_1^k\rangle}{|L_1|^2} + \frac{|a_2^k\rangle}{|L_2|^2} + \dots = \sum_w \frac{|a_w^k\rangle}{|L_w|^2} = \sum_{a_w^k=00\dots0}^{\overbrace{11\dots1}^{v}} L_{a_w^k} |a_w^k\rangle \quad (23)$$

Where, $L_{a_w^k}$ satisfies equation (21).

C. QRL based State Value Updating

In QRL, every possible state $|S^k\rangle = \{|s_1^k\rangle, |s_2^k\rangle, |s_3^k\rangle \dots \dots |s_l^k\rangle \dots .\}$ transformed into orthogonal Eigen states by unitary transformation $\left|s_l^{k,j}\right\rangle : |S^k\rangle = \sum_{l=1}^{N_{st}} L_l \left|s_l^{k,j}\right\rangle$. By applying quantum parallelism to u qubit states, which gives 2^u number of states and their state-values are parallelly updated according to one-step TD(0) updating rule.

$$V(s_l^{k,j}) \leftarrow V(s_l^{k,j}) + \zeta(U^{k,j} + \gamma V(s_l^{k,j+1}) - V(s_l^{k,j})) \quad (24)$$

For u qubit, there are 2^u exponential computation spaces required for simultaneous state-values updating operation improving the speed of learning over traditional RL.

D. Updating of Probability Amplitude

The selection of an Eigen action $|a_w^{k,j}\rangle$ related to superposition action $\left|a_w^{k,j^{N_{ac}}}(s_l^{k,j})\right\rangle$ depends upon their occurrence probability $\left|L_{a_w^k}\right|^2$ according to collapse postulate. After taking an action, reward of system defines the action property either "good" or "bad" actions. The probability amplitude of an action is updated (amplify or shrink) according to corresponding reward. Therefore, updating the amplitude of an action is the key to exploration (trial-and-error) and exploitation (experience) strategy for QRL agent. For v qubit action states, the Eigen actions in superposition state consist of 2^v actions. Choosing an action $|a_w^{k,j}\rangle$ is directly related to updating its probability amplitude based on Grover's iteration. First of all, assign the probability amplitude of all possible Eigen actions 2^v of v qubits with same weight. Applying Hadamard transformation ($H^{\otimes v}$) on each v qubit with initial state ($|0\rangle^{\otimes v}$=$|0\rangle$) assigns equal probability weight to each action.

$$H^{\otimes v} \overbrace{|00\dots0\rangle}^{v} = \frac{1}{\sqrt{2^v}} \left(\sum_{a_w=00\dots0}^{\overbrace{11\dots1}^{v}} |\, a_w^{k,j}\rangle\right) = |a_0^{k,j^{(v)}}\rangle \quad (25)$$

The probability amplitude of selecting an Eigen action $|\, a_w^{k,j}\rangle$ irrespective of its value $a_w^{k,j}$ is given as

$$\left|\langle\, a_w^{k,j} \left|a_0^{k,j^{(v)}}\right\rangle\right| = \frac{1}{\sqrt{2^v}} \quad (26)$$

Grover's iteration is used to reinforce the probability amplitude of good action based on obtained reward. The Grover's iteration is composed of two unitary reflections. First reflection corresponds to oracle transformation ($U_{a_w^{k,j}}$), which negate the amplitude of selected Eigen action.

$$U_{a_w^{k,j}} = I - 2|\, a_w^{k,j}\rangle\langle\, a_w^{k,j}| \quad (27)$$

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

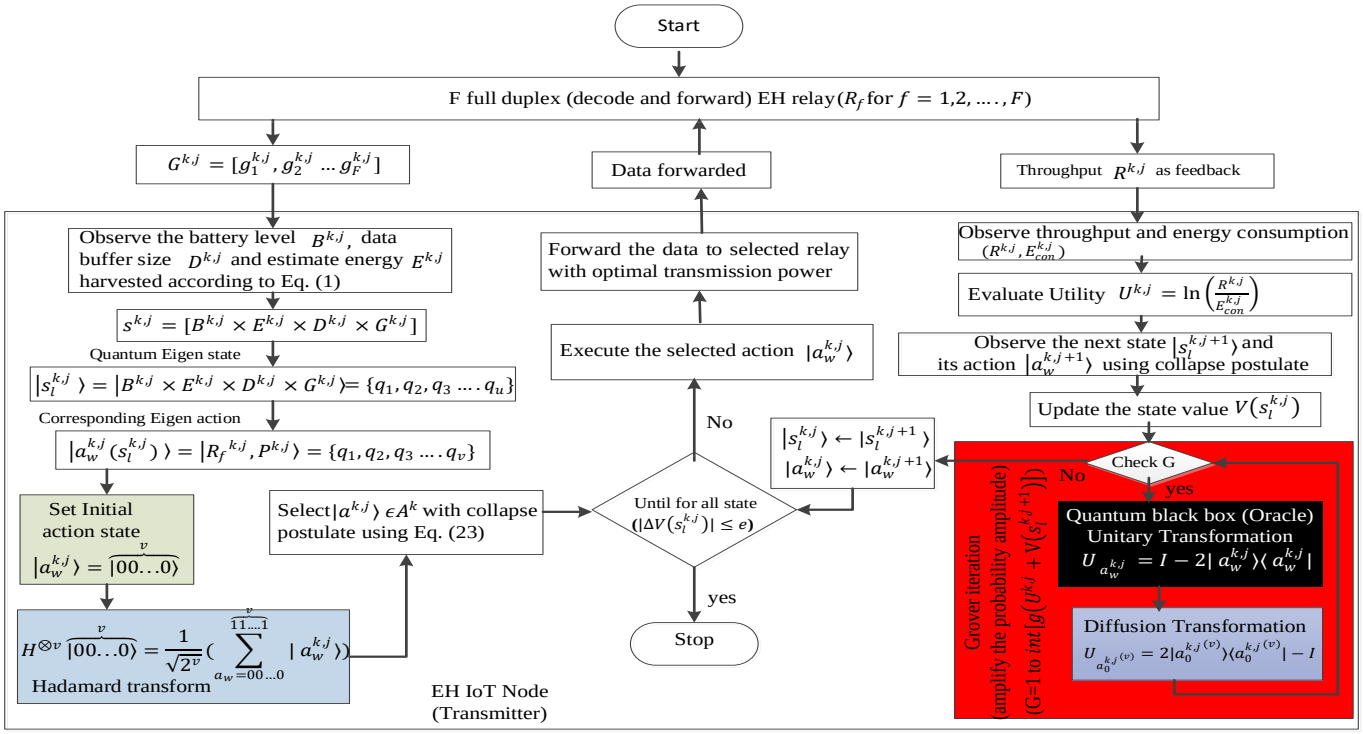

Fig. 4. Flowchart of the proposed QRL-RPS algorithm

The second reflection corresponds to diffusion transformation ($U_{a_0^{k,j}(v)}$), which enhances the amplitude of good action and suppresses the amplitude of other actions.

$$U_{a_0^{k,j}(v)} = 2|a_0^{k,j}(v)\rangle\langle a_0^{k,j}(v)| - I \tag{28}$$

where, I denotes the unitary matrix with suitable dimension. Thus, Grover transformation is represented as

$$U_{Gr} = U_{a_0^{k,j}(v)} U_{a_w^{k,j}} \tag{29}$$

The repeated number of Grover's transformation U_{Gr} applies on action ($a_0^{k,j}(v)$) and amplifies the probability amplitude of good action. When an action $a_w^{k,j}$ is carried out in Eigen state$|\ s_l^{k,j}\ \rangle$ respective reward is obtained. Then, according to the obtained reward, probability amplitude of good action is amplified by iterating G times of Grover's transformation. The number of iteration can be calculated as $G = int[g(U^{k,j} + V(s_l^{k,j+1})]$, where g is proportionality constant. After each updating, probability amplitude normalizes to $\sum_{a_w^k=00\ldots0}^{\widehat{11\ldots1}^v} \left|L_{a_w^k}\right|^2 = 1$. Thus, the optimal policy for joint relay and transmit power (action) can be constructed by simply selecting an action with higher amplitude value in each state using quantum collapse postulates.

VI. QRL BASED JOINT OPTIMAL RELAY AND TRANSMIT POWER SELECTION (QRL-RPS)

Initially, we establish a relationship between action (state) of RL and Eigen action (Eigen state) of QRL by choosing an observable of quantum system. Observable provides a set of Eigen vectors of dimension 2^v actions (for 2^u Eigen states), that form a complete set of orthonormal bases in a Hilbert space. When quantum superposition action ($a_0^{k,j}(v)$) is observed, an Eigen action $a_w^{k,j}$ is obtained. The action is selected according to collapse postulate using Eq. (23). After execution of obtained Eigen action $a_w^{k,j}$, node state changes to next state $s_l^{k,j+1}$ with its state value $V(s_l^{k,j+1})$ and corresponding reward $U^{k,j}$ during t_j. To update state value $V(s_l^{k,j})$ one-step TD (0) rule is used. The reward $U^{k,j}$ and state value $V(s_l^{k,j+1})$ are used to determine the value of G for amplification of probability amplitude of good action. The learning process repeats to small positive number e such that $|\Delta V(s_l^{k,j})| \leq e$, where $\Delta V(s_l^{k,j})$ is the difference between previous and current state-value and $e = 10^{-3}$. For better understanding of workflow of proposed framework, a flowchart is given in Fig. 4. The quantum black box (oracle ( )) modify system state using phase shift by π radian if the system is already into correct state or do nothing. Simply, oracle function negates probability amplitude of system and leaves the system in correct state [27]. Further, an algorithm is also presented (Algorithm-1) to show the working in terms of starting from MDP and their solution using quantum learning.

Algorithm-1: QRL-RPS

1. **Initialization:**
   - I. Initialize the value of ζ, γ, g, e and time slot j;
   - II. For a node $N_k(j)$, observe the Quantum Eigen state $|s_l^{k,j\,u}\rangle = \sum_{s_l^k=00\ldots0}^{\widehat{11\ldots1}^u} L_{s_l^k}|s_l^{k,j}\rangle$ and its corresponding Eigen action $|a_w^{k,j\,v}(s_l^{k,j})\rangle = \sum_{a_w^k=00\ldots0}^{\widehat{11\ldots1}^v} L_{a_w^k}|a_w^{k,j}\rangle$
   - III. Initialize the state-value $V(s_l^{k,j})$ of each state $|s_l^{k,j}\rangle$ corresponding to its available Eigen actions $|a_w^{k,j\,v}(s_l^{k,j})\rangle$.
   - IV. Observe initial state $|s_l^{k,j}\rangle$ of node $N^k(j)$ and its corresponding all available Eigen actions $|a_w^{k,j}\rangle$.

V. Initialize all the action qubits (v) to zero using quantum identity gate for observed state $|s_l^{k,j}\rangle$.

VI. Apply Hadamard gate on each action qubits to provide equal amplitude value to each available actions of state $|s_l^{k,j}\rangle$.

VII. Select an action using quantum collapse postulate Eq.(23) and get $|a_w^{k,j}\rangle$.

2. Repeat (for each episode)

While ($|\Delta V(s_l^{k,j})| \leq e$)

I. Execute the action $|a_w^{k,j}\rangle$ i.e. relay node is selected and transmit power is allocated to N_k in state $|s_l^{k,j}\rangle$.

II. N_k transmits information to selected relay using allocated transmit power

III. Relay sends the calculated throughput $R^{k,j}$ to N_k as feedback

IV. N_k evaluates the utility $U^{k,j}$ using throughput $R^{k,j}$ and energy consumption $E_{con}^{k,j}$.

V. Observe the next state $|s_l^{k,j+1}\rangle$ and its corresponding action $|a_w^{k,j+1}\rangle$ using collapse postulate.

VI. Update the state value $V(s_l^{k,j})$ using Eq.(24)

/amplify the probability amplitude of good action using Grover's Iteration/

VII. **For** (G=1 to $int[g(U^{k,j} + V(s_l^{k,j+1})])$

$$U_{Gr}|a_w^{k,j}\rangle = U_{a_0^{k,j}(v)} U_{a_w^{k,j}} |a_w^{k,j}\rangle$$

VIII. Set $|s_l^{k,j}\rangle \leftarrow |s_l^{k,j+1}\rangle$ and $|a_w^{k,j}\rangle \leftarrow |a_w^{k,j+1}\rangle$

End While

A. Analysis of the Proposed QRL-RPS Algorithm

In this section, some conceptual characteristic of the proposed algorithm has been discussed. Three major properties are presented including: (1) an asymptotic analysis of convergence property, (2) time complexity and (3) feasibility study for QRL-RPS algorithm.

1. *Convergence Property*- For updating of state-value in QRL, TD(0) method is used. For the proposed learning problem, QRL algorithm converges to its optimal value $V^*(s_l^{k,j})$ by utilizing appropriate exploration strategy under the following condition-

$$\lim_{T\to\infty} \sum_{j=1}^{T} \zeta_j = \infty, \qquad \lim_{T\to\infty} \sum_{j=1}^{T} {\zeta_j}^2 < \infty \qquad (30)$$

In QRL, we use same one-step TD(0) updating rule as in traditional RL. So, the convergence property of the proposed QRL is same as traditional TD algorithm in RL [26].

2. *Time Complexity Analysis*- The time complexity of the proposed algorithm mainly depends upon updating of state-value $V(s_l^{k,j})$ and Grover's iteration in the line number VI and VII, respectively. The updating Eq. (24) for value iteration take time complexity of O($|N_{st} \times O(v)|$), for each updation of single state-value $V(s_l^{k,j})$, because it uses collapse postulate to measure superposition action-state (measure v number of qubits) and provide the action with maximum probability in constant time O(v). And the amplitude amplification of action is done by Grover's iteration (VII) in $O(|\sqrt{N_{ac}}|)$. The value update and Grover's amplitude amplification sweeps through state-space N_{st} in order to converge the algorithm till $(|\Delta V(s_l^{k,j})| \leq e)$ that includes another factor of N_{st}, making overall time complexity $O\left(N_{st} \times \left(|N_{st} \times O(v)| + |\sqrt{N_{ac}}|\right)\right)$ i.e. $O\left(N_{st} \times \left(|N_{st}| + |\sqrt{N_{ac}}|\right)\right)$.

3. *Feasibility Analysis*- The proposed algorithm mainly depends upon two operations such as initialization of equally weighted superposition state and Grover's iteration. As, Grover's search algorithm also uses above specified operations. And, these operations can be done by using quantum gates and realization of these gates (Hadamard and Identity) are possible through creating quantum virtual machines on classical machines using available software development kits such as Rigetti's Forest, Scikit-quantum python module. Thus, already implemented Grover's search algorithm proves feasibility of our proposed algorithm.

V. Simulation and Result Analysis

A. Simulation Environment Settings

The simulation of the proposed framework for network utility maximization is implemented using quantum virtual machine (QVM) hosted on Rigetti's Forest platform with supporting programming language: quantum instruction language (Quil) which requires pyQuil Python-package and Python 3.8.1. The major inbuilt functions of pyQuil that are included in simulation script of proposed framework, are as follows. 1) get_qc(): to get a simulated quantum computer on the local server machine. 2) local_forest_runtime(): to make sure both QVM and Quil complier are available. 3) qvm.run_and_measure(): run program; collapse the state with a measurement and return result. 4) declare(): to get stored result in quantum register. The major quantum gates used in simulation are I(): Identity gate, and H(): Hadamard gates. The MATLAB Simulink model of photovoltaic cell panel is used to implement the energy-harvesting model [28].

We consider an EH communication scenario consisting of a source node, 16 full-duplex relay nodes and a sink node that are randomly scattered over $100 \times 100\ m^2$ network area. The reasons behind consideration of full-duplex relay nodes over half-duplex relay nodes in results is two folds: (i) It doubles the throughput of network without scavenging any extra channel resource. (ii) Using separate antennas for receiving and transmitting signal can reduce self-interference in static relay makes it more spectrally efficient and delay-tolerant in practical channel conditions [29]. Thus, for using half-duplex relay node in network force us to allocate two orthogonal time or frequency channel for data communication. Which puts extra cost in network and throughput of network also degrades. The network system can be in any state from the considered eight states $|s^{k,j}\rangle$ and each state has corresponding $A^k(s^{k,j}) = \{|a_0^{k,j}\rangle, |a_1^{k,j}\rangle, \ldots, |a_{15}^{k,j}\rangle\}$ actions. Eight state and its corresponding 16 actions are represented by $u = 3$ and $v = 4$ qubits in QVM. The channel bandwidth of all communication links is considered $1000\ Hz$ and entire link experience Rayleigh flat fading. The MATLAB function is used to generate time-varying Rayleigh fading channel value $m^{k,j}$ and channel coefficient value $h^{k,j}$ [30], which is randomly generated while satisfying its statistical properties. The initial energy level of all sensors is considered $10\ mJ$.

B. Result Analysis

1) Convergence Performance Against Time Slots

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

Fig. 5-7. illustrates the convergence property of the proposed technique and the state-of-the-art techniques in terms of throughput, energy consumption and utility respectively against time slot t_j of the learning phase using parameter $\zeta = 0.05$ and $\gamma = 0.85$. We observed in the fig. 5, the throughput of the proposed technique increases as the time slot increases until $t_j = 180$ and after that, it converges to optimal value for each state of the system. However, for DQ-RSS and QL-RSA the technique converges to optimal throughput value for each state after $t_j = 410$ and $t_j = 580$ respectively. Additionally, from the fig.6, it is observed that energy consumption decreases with increase in time slot and converges after $t_j = 180, 410$ and 580 for QRL-RPS, DQ-RSS and QL-RSA respectively. Thus, the proposed technique learns the optimal policy faster and converges in less time, which in turn reduces the energy consumption for data transmission using different selected relay and transmit power that enhances the throughput as compared to the other techniques. Further, as the utility is computed in terms of throughput and energy consumption, so on combining the result of the fig. 5 and fig. 6 the utility mapped with time slot is shown in the fig. 7. It can be observed from the fig. 7, the proposed technique achieves higher utility $(U^{k,j})$ after $t_j = 180$ as time slot increases whereas DQ-RSS and QL-RSA lagging behind by 27 % and 40% respectively. Therefore, the proposed algorithm adaptively obtains optimal policy for joint relay and transmits power selection after its convergence in less time.

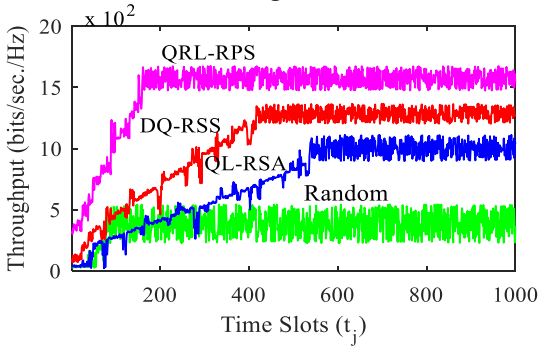

Fig. 5. Convergence performance comparison of throughput

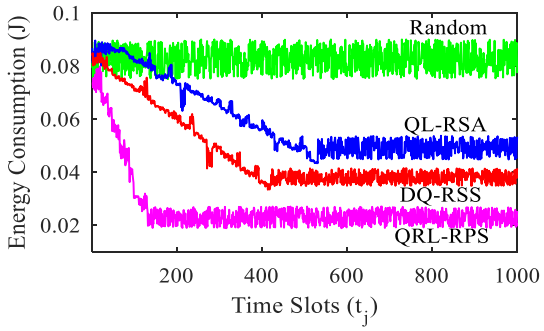

Fig. 6 .Convergence performance comparison of energy consumption

This is due to the fact that the proposed technique uses the concept of quantum information processing. The quantum processing utilizes the concept of state superposition and quantum parallelism. Further, for optimal relay and transmit power selection (action) it utilizes amplitude amplification of actions using Grover's iteration and quantum collapse postulates. However, QL-RSA and DQ-RSS uses random and ϵ-greedy policy respectively, which provides slow learning speed and is not preferable for complex problem having large number of state and action space. Furthermore, the random selection policy randomly selects any relay and transmits power which is not optimal and makes the system to consume more energy and have lower throughput that leads to less utility.

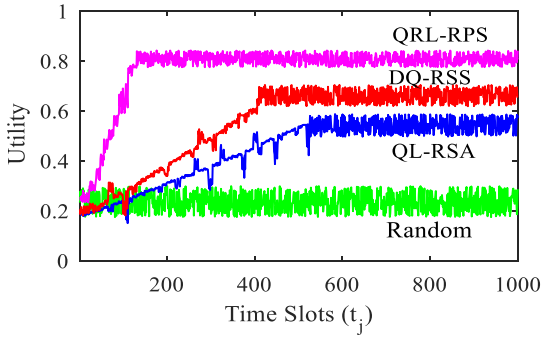

Fig. 7. Convergence performance comparison of utility

2) System Performance Against SNR

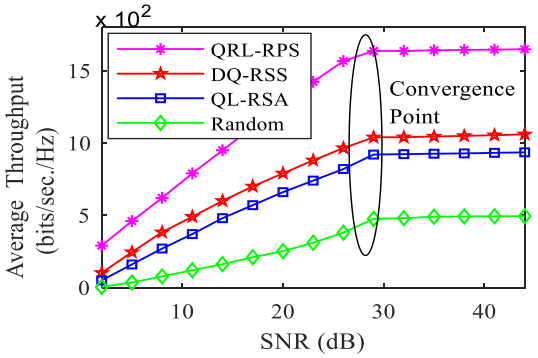

Fig. 8. Throughput as a function of SNR

A comparison of average throughput and average energy consumption between QRL-RPS and state-of-the-art techniques with different SNR value are presented in Fig. 8 and Fig. 9 respectively. The throughput and energy consumption of all considered techniques are a monotonically increasing function of SNR of the communication channel with parameters $\zeta = 0.05$ and $\gamma = 0.85$. This can be attributed to the reason that high SNR value indicates the signal is less distorted by channel noise and the possibility of successful and correct data decoding at the receiver side enhances, which as a result require more energy. It can be clearly observed that after a certain SNR value (29 dB) average throughput and average energy consumption becomes saturated is fourfold (i) self-interference cancellation capability of full-duplex relay node reaches up to maximum capability (ii) the limited computational capability of IoT node to decode the received signal; (iii) QRL-RPS has to fulfilled the constraint C_3 and C_4; such that throughput of relay nodes depends upon the data arrival from the previous EH-IoT node to defer the buffer overflow condition. (iv) and, selection of transmit power is bounded by constraint defined in C_5. It is evident from the results that QRL-RPS outperforms DQ-RSS and QL-RSA by improving the throughput and reducing the energy consumption. As, the utility enhances by increasing the throughput and lowering down the energy consumption. The results in Fig. 8 and Fig. 9 are in favor of enhancing the

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

average utility of the QRL-RPS technique as compared to DQ-RSS and QL-RSA as shown in Fig. 10.

The QRL-RPS technique achieves such good performance by taking the advantage of the quantum property such as quantum superposition and quantum parallelisms used for selection of joint optimal relay and transmit power at each network state. The QL-RSA and DQ-RSS use random and ϵ-greedy policy respectively. The problem with these schemes is that the ϵ-greedy policy selects equally among all available actions while exploring. In addition, it is difficult to select a proper value for ϵ, which can provide a balance between the exploration and exploitation strategy for action selection. Whereas, the QRL-RPS technique utilizes the probabilistic approach for optimal action selection, that is motivated by the collapse postulate of quantum measurement which do not require any parameter settings and overcome the problem of QL-RSA and DQ-RSS. Thus, the proposed technique makes the learning agent to learn an optimal policy faster in an uncertain environment as compared to the other techniques. The random technique shows worst performance as compared to the other techniques.

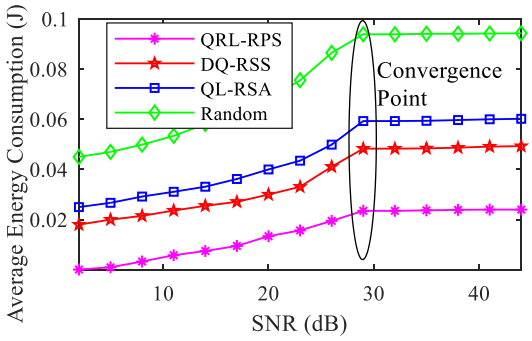

Fig. 9. Energy consumption as a function of SNR

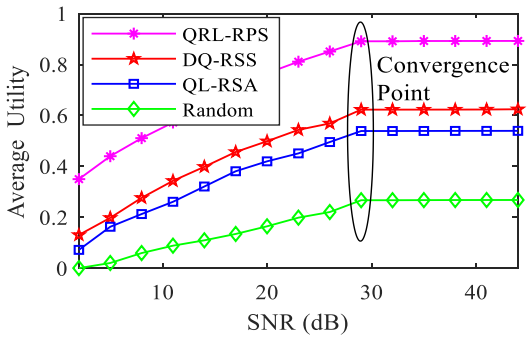

Fig. 10. Utility as a function of SNR

3) *Learning Performance of the Proposed Technique with Different Learning Rates*

The result in the Fig. 11 (a) shows the impact of varying learning rates i.e. $\zeta = \{0.005, 0.05, 0.5\}$ on QRL-RPS convergence rate. It can be clearly observed that for smaller value of learning rate (i.e.,$\zeta = 0.005$), the learning agent in initial trail of learning takes more than 1000 steps to learn optimal policy. This can be attributed to the reason that at initial trail learning agent behaves naive, does not have knowledge about environment, and spends more number of steps in exploring the environment. Further, as the number of learning trails increases, learning agent settle down between 380 to 440 steps to reach convergence. This is because the learning agent gains knowledge about environment and not always wastes the steps in exploration. It is worthy to note that for learning rate $\zeta = 0.05$**,** the proposed technique takes less number of steps for convergence as compared to $\zeta = 0.005$ and 0.5 . This observation affirms that learning agent is able to balance between exploration of environment and exploitation of knowledge. The reason behind is Grover's amplitude amplification method makes the learning agent smarter to learn the optimal policy faster while searching over all the available action space. Only for the initial trail of learning rate $\zeta = 0.05$, the learning agent takes 590 steps and thereafter for most of the upcoming trails learning agent learns the optimal policy at the cost of between 160 to 210 trials**.** It is also notable from the result that **for** $\zeta = 0.5$, the proposed technique is unable to converge to suitable value. This observation affirms that the learning agent frequently updates the old reward and is not able to learn the optimal policy. Whereas fig 11 (b) shows the learning performance of the presented technique in terms of taken steps until convergence regard to learning rate $\zeta \in \{0.02, 0.05, 0.1\}$ is about 230. It provide guarantees to faster and better selection of the joint transmit power and relay for the dynamic environment attributed as robust technique.

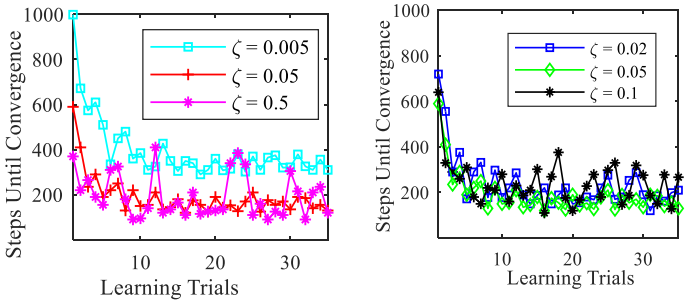

Fig. 11 (a-b) Convergence of the QRL-RPS technique with ζ

4) *Convergence of the Proposed Technique with Different Discount Factor and Learning Rate*

The result in Fig. 12 shows the joint impact of varying discount factor ($0.1 \leq \gamma \leq 1$) and learning rate ($0.01 \leq \zeta \leq 0.1$) on average utility of QRL-RPS technique. It can be clearly observed that QRL-RPS experiences higher utility (0.86) for trained optimal policy at $\gamma = 0.85$ and $\zeta = 0.05$. This observation affirms that the achieved reward in recent time slots has more impact on utility for the current time slot rather than the reward achieved in longer time horizon. It is evident from the result that for $\gamma = 1$**,** and $\zeta =$**0.1,** the learning agent gives equal chance to all the previous reward under unbalanced exploration and exploitation strategy, respectively. Thus, it contains most irrelevant reward of longer time horizons at the cost of large fluctuations, which have less impact in calculation of long-term expected utility. However, when γ value is too small, i.e., in range {0.1, 0.5}, and learning rate is set to any value, the trained optimal policy pays more attention to the most immediate reward rather than future rewards and utility converges to bad value. It is worthy to note that for the lower value of $\gamma = 1 - 1 * e^{-10}$, the utility of the proposed technique plunges to much lower value. This can be attributed to the reason that very long time old rewards cannot estimate the network behavior perfectly in terms of utility. It is also important to note that, QRL-RPS keeps good

Cite Aa:
Jaiswal, A., Kumar, S., Kaiwartya, O., Kashyap, P.K., Kanjo, E., Kumar, N. and Song, H., 2021. Quantum Learning Enabled Green Communication for Next Generation Wireless Systems. *IEEE Transactions on Green Communications and Networking*, pp. 1-14, 2021.

learning performance in the range of $\gamma \epsilon \{0.65,\ 0.9\}$ and $\zeta \epsilon \{0.05,\ 0.1\}$,and maximize the network utility above 0.73. Thus, the QRL-PRS is much more robust as learning performance is in control manner for wider range of learning rate and discount factor. This can be attributed to much more practical learning environment such as joint optimal relay and transmit power selection.

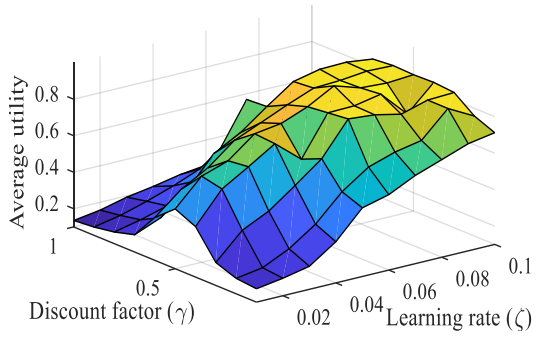

Fig. 12. Utility as a function of ζ & γ

5) *Impact of Data Buffer Size on Average Utility*

A comparison of average utility between QRL-RPS and state-of the-art techniques with varying data buffer size factor is presented in Fig. 13. For all considered techniques, the buffer size at relay node N_2 and N_3 is taken $D^{max,2} = \mu R^{1,j}(B^{max,2})$ and $D^{max,3} = \mu R^{2,j}(B^{max,3})$ respectively, where μ denotes a tunable parameter, and $R^{1,j}$and $R^{2,j}$ are expected throughput of N_2 and N_3 respectively. The selection of optimal relay and transmit power depends upon current buffer level of nodes N_2 and N_3 in each time slot. It is evident from the result that QRL-RPS technique handles the data buffer overflow condition better, and its average utility is higher as compared to other techniques for the considered range of μ. This can be attributed to the reason that QRL-RPS uses the concepts of amplitude amplification and collapse postulates of quantum mechanics for action selection, which handles the overflow condition better than state-of-the-art techniques. Further, it makes the learning agent to better explore the available options of relays and transmit powers in order to increase the long-term average utility in an uncertain environment with only casual knowledge (buffer level) available at the transmitter. DQ-RSS and QL-RSA techniques use ϵ-greedy and random policy for action selection respectively, whose exploration and exploitation strategy is not so good. And, the selected relay and transmit power of node N_2 and N_3 are not able to handle the buffer overflow condition properly. The average utility of random technique is the lowest, as it does not have any learning agent and it selects the relay and transmit power randomly for data transmission.

Further, it can be observed that all the considered techniques achieve less utility using small value of μ because data received from N_1 and N_2 are not completely stored in the data buffer of N_2 and N_3 respectively and are dropped. As, this scenario increases the probability of data retransmission and further increases the total energy consumption. As evident from the result that the average utility saturates and data buffer overflow conditions happen less frequently for all the techniques at approx. value of $\mu = 2.7$ when the data buffer size is more at N_2 and N_3 as compared to the data rate received from N_1 and N_2 respectively. However, for large data buffer size (large value of μ) as compared to $R^{1,j}$ and $R^{2,j}$ for N_2 and N_3 respectively, its effect on the network performance is minimum. Because there is limitation on data size arrival at N_2 and N_3 (because of limited power available at N_1 and N_2 respectively) which can be stored in proper buffer size for $\mu = 3$, Therefore, larger value of buffer size does not affect much the network performance in terms of utility.

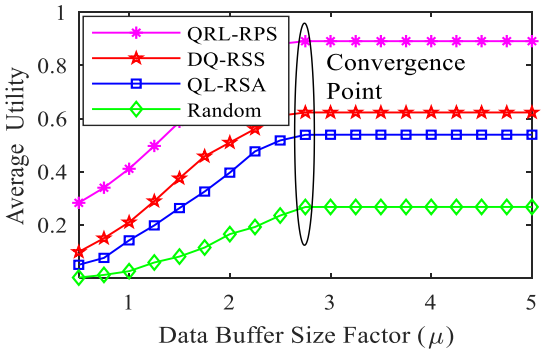

Fig. 13. Utility as a function of data buffer size factor (μ)

6) *Grover's Amplitude Value and Occurrence Probability of Different Actions*

The result in Fig. 14 shows the amplitude value and occurrence probability for different actions of QRL-RPS learning agent after applying Grover's transformation. The heart of the proposed QRL-RPS technique for joint optimal relay and transmit power selection (action) is its amplitude amplification method, and it depends upon Grover's iteration process. It helps the learning agent to take better decision for action selection in each time slot in order to learn the optimal policy. The aforementioned method amplifies the amplitude of "good action" in each Grover's iteration.

For the simulation, single state $|s_3^{1,100}\rangle$ of the source node N_1 and its corresponding 16 actions (4 qubit representations) in time slot $t_j = 100$ with parameters $\zeta = 0.05$, $\gamma = 0.85$ and $g = 3$ are considered. The superposition action for state, $|s_3^{1,100}\rangle = \{0.08J, 0.025J, 12 * 10^3 bits, [0.062, 0.047,\ 0.074, 0.616, 0.097, 0.0357, 0.0017, 0.0253]\}$ of N_1 in $t_j = 100$ after amplitude amplification is computed as

$$\begin{aligned}\left|a_{w_{0..15}}^{1,100\ 4}(s_3^{1,100})\right\rangle = {} & (0.1+0j)|0000\rangle + (0.1+0j)|0001\rangle \\ & + (0.1+0j)|0010\rangle + (\mathbf{0.86+0j})|\mathbf{0011}\rangle \\ & + (0.1+0j)|0100\rangle + (0.1+0j)|0101\rangle \\ & + (0.1+0j)|0110\rangle + (0.1+0j)|0111\rangle \\ & + (0.1+0j)|1000\rangle + (0.33+0j)|1001\rangle \\ & + (0.1+0j)|1010\rangle + (0.1+0j)|1011\rangle \\ & + (0.1+0j)|1100\rangle + (0.1+0j)|1101\rangle \\ & + (0.1+0j)|1110\rangle + (0.1+0j)|1111\rangle\end{aligned}$$

where, sum of the occurrence probability of action after amplitude amplification is equal to 1. It is evident from the results that the action collapse phenomenon collapses the superposition action $\left|a_{w_{0..15}}^{1,100\ 4}(s_3^{1,100})\right\rangle$ to action $|0011\rangle$ ($\left|{R_f}^{k,j}, P^{k,j}\right\rangle = |r_4, 0.8\ dBm\rangle$) with amplitude $(0.86 + 0j)$ and occurrence probability $|(0.86 + 0j)|^2 = 0.75$.

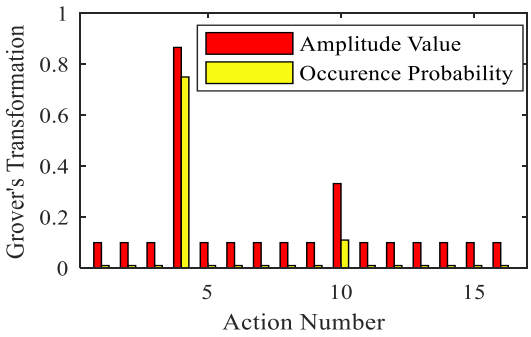

Fig. 14. Grover's transformation versus action number

VII. Conclusions and Future Scope

In this paper, a novel QRL-based framework is proposed to inspect the problem of joint optimal relay and transmit power selection for network utility maximization in energy-constrained sensor enabled-IoT. The states of the environment and actions of the corresponding state are represented by qubits using the concept of quantum superposition and action selection is motivated by the collapse phenomenon of quantum computing. A QRL-RPS algorithm is proposed to provide optimal policy for the formulated selection problem. The simulation results show that the proposed technique performs better as compare to state-of-the-art techniques in terms of convergence speed and network utility by 40% and 27% against QL-RSA and DQ-RSS respectively. Our proposed idea not only enhances the working of the current learning algorithm on conventional computers, but also encourages the evolution of related research fields like machine learning and quantum computation. In our future research, the team will investigate more use cases for quantum learning enabled green computing in next generation networks such as in connected traffic networks, big data processing and visualization [31, 32]. Integrating traditional optimization approach with quantum learning for green communication will also be the quest [33].

References

[1]. N. Kumar, R. Chaudhry, O. Kaiwartya, N. Kumar and S. H. Ahmed, "Green Computing in Software Defined Social Internet of Vehicles," in IEEE Transactions on Intelligent Transportation Systems, pp-1-10, 2020. (online published) doi: 10.1109/TITS.2020.3028695.

[2]. Farhan, L., Kharel, R., Kaiwartya, O., Quiroz-Castellanos, M., Alissa, A. and Abdulsalam, M., 2018, July. A concise review on Internet of Things (IoT)-problems, challenges and opportunities. In 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP) (pp. 1-6), IEEE, Budapest, Hungary, 18-20 July 2018

[3]. Asif-Ur-Rahman, M., Afsana, F., Mahmud, M., Kaiser, M.S., Ahmed, M.R., Kaiwartya, O. and James-Taylor, A, "Toward a Heterogeneous Mist, Fog, and Cloud-Based Framework for the Internet of Healthcare Things," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4049-4062, June 2019, doi: 10.1109/JIOT.2018.2876088.

[4]. S. Kumar, O. Kaiwartya, M. Rathee, N. Kumar and J. Lloret, "Toward Energy-Oriented Optimization for Green Communication in Sensor Enabled IoT Environments," in IEEE Systems J., vol.(14)4, pp. 4663-4673, Dec. 2020.

[5]. Kumar, K.; Kumar, S.; Kaiwartya, O.; Cao, Y.; Lloret, J.; Aslam, N. Cross-Layer Energy Optimization for IoT Environments: Technical Advances and Opportunities. Energies, 10(1), 2073, 2017.

[6]. A. Jaiswal, S. Kumar, O. Kaiwartya, M. Prasad, N. Kumar, & H. Song, "Green computing in IoT: Time slotted simultaneous wireless information and power transfer," Computer Communications, 168, pp. 155-169, 2021.

[7]. Khatri, A., Kumar, S., Kaiwartya, O., Aslam, N., Meena, N. and Abdullah, A.H., Towards green computing in wireless sensor networks: Controlled mobility–aided balanced tree approach. International Journal of Communication Systems, 31(7), p.e3463, 2018..

[8]. L. Farhan, O. Kaiwartya, L. Alzubaidi, W. Gheth, E. Dimla and R. Kharel, "Toward Interference Aware IoT Framework: Energy and Geo-Location-Based-Modeling," in IEEE Access, vol. 7, pp. 56617-56630, 2019.

[9]. A. Jaiswal, S. Kumar, O. Kaiwartya, N. Kumar, H. Song and J. Lloret, "Secrecy Rate Maximization in Virtual-MIMO Enabled SWIPT for 5G Centric IoT Applications," in IEEE Systems Journal, doi: 10.1109/JSYST.2020.3036417.

[10]. V. D. Valerio, F. L. Presti, C. Petrioli, L. Picari, Spaccini, & S. Basagni, "CARMA: Channel-aware reinforcement learning-based multi-path adaptive routing for underwater wireless sensor networks," IEEE Journal on Selected Areas in Communications, 37(11), pp. 2634-2647, 2019.

[11]. M. A. Jadoon, & S. Kim, "Relay selection Algorithm for wireless cooperative networks: a learning-based approach," IEEE Journal. Selected. Areas Communication, vol. 11, no. 7, 1061-1066, 2017.

[12]. Y. Su, X. Lu, Y. Zhao, L. Huang, & X. Du, "Cooperative Communications with Relay Selection Based on Deep Reinforcement Learning in Wireless Sensor Networks," IEEE Sensors J., 19(20), pp. 9561-9569, 2019.

[13]. S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, M. Asaduzzaman, "Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future," in IEEE Access, 7, 46317-46350, 2019.

[14]. Zhang, C., Patras, P. and Haddadi, H., "Deep learning in mobile and wireless networking: A survey", IEEE Communications Surveys & Tutorials, 21(3), pp.2224-2287, 2019.

[15]. O. J. Pandey and R. M. Hegde, "Low-Latency and Energy-Balanced Data Transmission Over Cognitive Small World WSN," in IEEE Trans. On Vehicular Technology, Vol. (67), no. (8), pp. 7719-7733, Aug. 2018.

[16]. A.S. Ibrahim, A. K. Sadek, W. Su, & K. R. Liu, "Cooperative communications with relay-selection: when to cooperate and whom to cooperate with?," IEEE Trans. on wireless comm., vol. 7, no. 7, pp. 2814-2827, 2008.

[17]. Y. Zhu, Z. Li, D. Sui, D. Li, J. Kong, & F. Hu, "Group-Based Relay Selection in Simultaneous Wireless Information and Power Transfer Network," IEEE Access, vol. 6, no. 1, pp. 49019-49028, 2018.

[18]. M. K. Shukla, H. H. Nguyen and O. J. Pandey, "Multiuser Full-Duplex IoT Networks With Wireless-Powered Relaying: Performance Analysis and Energy Efficiency Optimization," in IEEE Transaction on Green Communication and Networking, vol. 4, no. 4, pp.982-997, Dec.-2020.

[19]. Y. Wu, L. ping Qian, Huang & X. Shen, "Optimal relay selection and power control for energy-harvesting wireless relay networks," IEEE Trans. on Green Communications and Networking, 2(2), pp. 471-481, 2017.

[20]. I. Ahmed, A. Ikhlef, R. Schober, & R. K. Mallik, "Joint power allocation and relay selection in energy harvesting AF relay systems. IEEE Wireless Communications Letters, vol. 2, no. 2, pp. 239-242, 2013.

[21]. Y. Xiao, Z. Han, D. Niyato, & C. Yuen, "Bayesian reinforcement learning for energy harvesting communication systems with uncertainty," IEEE International Conference on Comm., pp. 5398-5403, 2015.

[22]. K. K. Nguyen, T. Q. Duong, N. A. Vien, N. A. Le-Khac, & M. N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," IEEE Access, vol. 7, pp. 100480-100490, 2019.

[23]. E. Strubell, "An introduction to quantum algorithms," COS498 Chawathe Spring, vol.13, pp. 1-35, 2011.

[24]. D. Dong, C. Chen, J. Chu, & T. J. Tarn, "Robust quantum-inspired reinforcement learning for robot navigation. IEEE/ASME transactions on mechatronics," vol. 17, no. 1, pp. 86-97, 2010.

[25]. C. L. Chen, D. Y. Dong, & Z.H. Chen, "Quantum computation for action selection using reinforcement learning," International Journal of Quantum Information, vol.4, no. 6, pp. 1071-1083, 2006.

[26]. R. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction. Cambridge," MA: MIT Press, 1998.

[27]. D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[28]. MATLAB/Simulink Model of Photovoltaic Cell, (https://www.github.com/motahhir/MATLAB-Simulink-Model-of-Photovoltaic-Cell-Panel-and-Array-), GitHub. Retrieved March 7, 2020.
[29]. T. Riihonen, S. Werner and R. Wichman, "Comparison of Full-Duplex and Half-Duplex Modes with a Fixed Amplify-and-Forward Relay," 2009 IEEE Wireless Comm. and Net. Conference, pp. 1-5, Budapest, 2009.
[30]. K. E. Baddour, N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," IEEE Transactions on Wireless Communications, vol. 4, no. 4, pp. 1650-1662, 2005.
[31]. Kaiwartya, O. and Kumar, S., 2014. Enhanced caching for geocast routing in vehicular Ad Hoc network. In Intelligent computing, networking, and informatics (pp. 213-220). Springer, New Delhi.
[32]. Prasad, M., Liu, Y.T., Li, D.L., Lin, C.T., Shah, R.R. and Kaiwartya, O.P.,. A new mechanism for data visualization with TSK-type preprocessed collaborative fuzzy rule based system. Journal of Artificial Intelligence and Soft Computing Research, 7(1), pp.33-46, 2017.
[33]. Kaiwartya, O. and Kumar, S., 2014, May. Geocasting in vehicular adhoc networks using particle swarm optimization. In Proceedings of the international conference on information systems and design of communication (pp. 62-66), Portugal, May 2014.

ANKITA JAISWAL is currently a Ph.D. research scholar at School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India. She received her M. Tech. degree in Computer Science and Technology from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2016 and B. Tech. degree in Computer Science and Engineering from Uttar Pradesh Technical University, India in 2013. Her research interests include 5G centric Wireless Sensor Networks and Internet of Things.

SUSHIL KUMAR (M'11 SM'18) is currently working as Assistant Professor at School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India. He received his Ph.D. degree in Computer Science from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2014. His research interest includes the area of vehicular cyber physical systems, Internet of things and wireless sensor networks. He is supervised/supervising many doctoral theses in vehicular communication, energy efficiency of terrestrial sensor networks, and green and secure computing in Internet of Things. He has authored and coauthored over 70 technical papers in international journals and conferences. He served as session chair in many international conferences and workshops. He is a reviewer in many IEEE/IET and other reputed SCI journals.

OMPRAKASH KAIWARTYA (M'14, SM'19) received the Ph.D. degree in computer science from Jawaharlal Nehru University, New Delhi, India, in 2015. He is currently a Senior Lecturer with the Department of Computer Science, Nottingham Trent University, Nottingham, U.K. He was previously a Research Associate with Northumbria University, Newcastle upon Tyne, U.K., in 2017 and a Postdoctoral Research Fellow with the University of Technology Malaysia, Johor Bahru, Malaysia, in 2016. His research interests include drone enabled networking, E-mobility centric electric vehicles, Internet of Things enabled smart services, connected vehicles, and next generation wireless systems. He is the Fellow with Higher Education Academy, U.K. He is also a Professional Member with British Computer Society, U.K. He is an Associate Editor and/or Guest Editor for the IEEE INTERNET OF THINGS JOURNAL, IEEE ACCESS, IET Intelligent Transport Systems, EURASIP Journal on Wireless Communication and Networking, MDPI Sensors, and Electronics.

PANKAJ KUMAR KASHYAP is working as District Informatics Officer in National Informatics Centre, Jammu & Kashmir (Union Territory) under Ministry of Electronics and Information Technology, Government of India, New Delhi. He received his Ph.D. degree and M.Tech degree in Computer Science and technology from Jawaharlal Nehru University, New Delhi in 2020 and 2014 respectively. His research area interest is load balancing and energy optimization in wireless sensor networks, Internet of Things or vehicles using machine learning approaches. Dr. Kashyap has published more than 5 papers in international journals and over 10 papers in international conferences. He is also published 3 book chapter in the lectures series of springer conferences. Dr. Kashyap is currently working on techniques quantum learning, deep learning, reinforcement learning, blockchain for energy optimization, energy trading, trust, security and privacy analysis in the field of agriculture, health and edge computing cloud based next generation IoT networks.

EIMAN KANJO is an Associate Professor in Pervasive Computing at Nottingham Trent University. She conducts research in Mobile Sensing, Pervasive Computing, Affective Computing and Data Science and she currently leads the Smart Sensing Lab and she also leads the technical development of the Smart Campus project. Dr Kanjo has worked previously at the University of Cambridge, Mixed Reality Laboratory, University of Nottingham and the International Centre for Computer Games and Virtual Entertainment, Dundee. Eiman Kanjo is a technologist who views the world through lens of pervasive sensors. She coined the phrase Mobile Sensing and written some of the earliest papers on the subject. She collaborates closely with industry, local authorities and end user organisations to solve real-world problems. She works closely with Nottingham County Council and Nottingham City Council on developing its Smart Cities strategy and she is the technical lead of the NTU Smart Campus project.

NEERAJ KUMAR (M'16, SM'17) received the Ph.D. degree in computer science and engineering from Shri Mata Vaishno Devi University, Katra (J&K), India, in 2009. He was a Post-Doctoral Research Fellow at Coventry University, Coventry, U.K. He is currently an Associate Professor with the Department of Computer Science and Engineering, Thapar University, Patiala, India. He has authored more than 200 technical research papers published in leading journals and conferences from the IEEE, Elsevier, Springer, John Wiley, etc. He is in the Associate Editor in IEEE Communications Magazine, IEEE Networks, IEEE Transactions on Sustainable Computing, Journal of Network and Computer Applications (Elsevier) and International Journal of Communication Systems (Wiley). He is one of highly cited authors in Web of Science in 2019.

HOUBING SONG (M'12–SM'14) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012, and the M.S. degree in civil engineering from the University of Texas, El Paso, TX, in December 2006. In August 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, where he is currently an Assistant Professor and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). He served on the faculty of West Virginia University from August 2012 to August 2017. In 2007 he was an Engineering Research Associate with the Texas A&M Transportation Institute. He has served as an Associate Technical Editor for IEEE Communications Magazine and a Guest Editor for IEEE Journal on Selected Areas in Communications (J-SAC), IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics and IEEE Network. He is the author of more than 100 articles. His research interests include cyber-physical systems, cybersecurity and privacy, internet of things, edge computing, big data analytics, unmanned aircraft systems, connected vehicle, smart and connected health, and wireless communications and networking. His research has been featured by popular news media outlets, including USA Today, U.S. News & World Report, Fox News, Forbes, WFTV, and New Atlas.